

Simultaneous Optimization of Both Node and Edge Conservation in Network Alignment via WAVE

Yihan Sun^{1,2,3}, Joseph Crawford¹, Jie Tang² and Tijana Milenković^{1,*}

¹Department of Computer Science and Engineering, Interdisciplinary Center for Network Science and Applications, and ECK Institute for Global Health, University of Notre Dame.

²Department of Computer Science and Technology, Tsinghua University.

³Computer Science Department, Carnegie Mellon University.

*Corresponding Author (E-mail: tmilenko@nd.edu)

Abstract. Network alignment can be used to transfer functional knowledge between conserved regions of different networks. Typically, existing methods use a node cost function (NCF) to compute similarity between nodes in different networks and an alignment strategy (AS) to find high-scoring alignments with respect to the total NCF over all aligned nodes (or node conservation). But, they then evaluate quality of their alignments via some other measure that is different than the node conservation measure used to guide the alignment construction process. Typically, one measures the amount of conserved edges, but only after alignments are produced. Hence, a recent attempt aimed to directly maximize the amount of conserved edges while constructing alignments, which improved alignment accuracy. Here, we aim to directly maximize both node and edge conservation during alignment construction to further improve alignment accuracy. For this, we design a novel measure of edge conservation that (unlike existing measures that treat each conserved edge the same) weighs each conserved edge so that edges with highly NCF-similar end nodes are favored. As a result, we introduce a novel AS, **Weighted Alignment VotEr** (WAVE), which can optimize any measures of node and edge conservation, and which can be used with any NCF or combination of multiple NCFs. Using WAVE on top of established state-of-the-art NCFs leads to superior alignments compared to the existing methods that optimize only node conservation or only edge conservation or that treat each conserved edge the same. And while we evaluate WAVE in the computational biology domain, it is easily applicable in any domain.

1 Introduction

1.1 Motivation

Network alignment aims to find topologically or functionally similar regions between different networks. It has applications in different areas, including computational biology^(25,35,29,24,42,37,44,14), ontology matching^(27,51,26,30), pattern recognition^(8,56), language processing⁽²⁾, social networks^(22,57), and others^(52,41,12,53,49). Our study focuses mainly on the computational biology domain.

Protein-protein interaction (PPI) networks have been the main focus of network alignment research among all biological networks. PPI network alignment can be used to transfer biological knowledge from the network of a poorly studied species to the network of a well studied species. This is of importance because not all cellular processes can easily be studied via biological experiments. For example, studying aging in human has to rely on across-species transfer of aging-related knowledge from model species⁽¹⁵⁾. And network alignment can be (and has been) used for this^(37,44,14). However, the problem is computationally intractable, as the underlying subgraph isomorphism problem is NP-complete⁽⁹⁾. Thus, network alignment methods are heuristics.

Network alignment can be local or global. Local network alignment aims to align well local network regions^(20,46,16,23,3,28,4,38,5). As such, it often fails to find large conserved regions between networks. Hence, majority of recent research has focused on global network alignment^(48,17,47,55,29,25,35,24,42,40,39,18,21,13,29,37,14,44,6,10), which can find large conserved regions between networks. Typically, global network alignment aims to generate one-to-one node mapping between

two networks⁽⁶⁾ (although exceptions exist that produce many-to-many node mappings or that align more than two networks⁽²⁹⁾, but such methods are out of the scope of our study).

Of one-to-one global network alignment methods, many consist of two algorithmic components, namely, a node cost function (NCF) and an alignment strategy (AS)^(37,14,10). NCF captures pairwise similarities between nodes in different networks, and AS then searches for good alignments based on the NCF information. It has already been recognized that when two methods of this two-component NCF-AS type are compared, to fairly evaluate the methods, one should mix and match the different methods’ NCFs and ASs, because NCF of one method and AS of another method could lead to a new method that is actually superior to the original methods^(37,14,10).

We base our work on established state-of-the-art NCFs of existing methods. Then, we propose a novel AS, **Weighted Alignment VotEr** (WAVE), which when used on top of the established NCFs leads to a new superior method for global network alignment. And while we evaluate our new method in the computational biology domain, the method is easily applicable in any domain.

1.2 Related Work

We focus on NCFs of two popular existing methods, MI-GRAAL⁽²⁴⁾ and GHOST⁽⁴²⁾, and we aim to improve with our new WAVE AS upon these methods’ ASs.

MI-GRAAL improves upon its predecessors, GRAAL⁽²⁵⁾ and H-GRAAL⁽³⁵⁾, by using the same NCF but by combining their ASs (see below). MI-GRAAL’s NCF computes topological similarity between extended network neighborhoods of two nodes^(32,50,31,34). It does so by relying on the concept of small induced subgraphs called graphlets (e.g., a triangle or a square)^(43,33), which are used to summarize the topology of up to 4-deep network neighborhood of a node into its graphlet degree vector (GDV)^(32,19,36). Then, GDV-similarity is used as MI-GRAAL’s NCF, which compares nodes’ GDVs to compute their topological similarity. MI-GRAAL also allows for integration of other node similarity measures into its NCF, such as protein sequence similarity. We recently showed^(37,14) that MI-GRAAL’s NCF is superior to another, Google PageRank algorithm-based NCF, which is used by IsoRank⁽⁴⁸⁾ and IsoRankN⁽²⁹⁾. Regarding AS⁽¹⁰⁾, MI-GRAAL’s AS combines GRAAL’s greedy seed-and-extend AS with H-GRAAL’s optimal AS that uses the Hungarian algorithm to solve linear assignment problem of maximizing total NCF over all aligned nodes.

GHOST’s NCF is conceptually similar to MI-GRAAL’s, as it also assumes two nodes from different networks to be similar if their neighborhoods are similar. However, the mathematical and implementation details of the two NCFs are different. Namely, GHOST’s NCF takes into account a node’s k -hop neighborhood, (in this study, $k = 4$). Then, its NCF computes topological distance (or equivalently, similarity) between two nodes by comparing the nodes’ “spectral signatures”. We recently fairly compared MI-GRAAL’s GDV-similarity-based NCF with GHOST’s “spectral signature”-based NCF within our above mix-and-match framework, concluding that MI-GRAAL’s NCF is superior or comparable to GHOST’s NCF, depending on data⁽¹⁰⁾. Hence, since none of the two NCFs was dominant in all cases, we consider both NCFs in our study. Just as MI-GRAAL, GHOST also allows for integration of protein sequence information into its NCF. Regarding AS, GHOST is also a seed-and-extend algorithm, like MI-GRAAL. However, GHOST’s AS considers quadratic (instead of linear) assignment problem. When we evaluated the two ASs, their performance was data-dependent⁽¹⁰⁾. Hence, we consider both ASs in our study.

There exist additional more recent network alignment methods⁽⁶⁾, both those that also belong to the category of NCF-AS methods, such as NETAL⁽⁴⁰⁾, and those that do not, such as MAGNA⁽⁴⁴⁾. These methods became available close to completion of our study, and as such, we were not able to include them into the design of our new method. (Hence, NETAL implements a different NCF compared to NCFs of MI-GRAAL and GHOST, along with a different AS compared to ASs of MI-GRAAL, GHOST, and WAVE.) However, we still consider these methods in our evaluation. Importantly, our goal is to show that when we use under an existing NCF (such as MI-GRAAL’s or GHOST’s) our new WAVE AS, we get alignments of higher quality compared to when using an

existing AS (such as MI-GRAAL’s or GHOST’s) on the same NCF. This would be sufficient to illustrate the superiority of WAVE. If in the process we also improve upon the more recent methods, such as those that use a different NCF and especially those that do not belong to the NCF-AS category, that would further demonstrate WAVE’s superiority.

1.3 Our Contributions and Significance

We introduce WAVE, a novel, general, and as we will show superior AS, which can be combined with any NCF. WAVE is applicable to any domain. We evaluate it on biological networks.

Its novelty and significance is as follows. The existing ASs use NCF scores to rapidly identify from possible alignments the high-scoring alignments with respect to the overall NCF (henceforth also referred to as node conservation). But, their alignment accuracy is then evaluated with some other measure that is different than NCF used to construct the alignments⁽⁴⁴⁾. Typically, one measures the amount of conserved (i.e., aligned) edges. Hence, a recent attempt aimed to directly maximize edge conservation during alignment construction⁽⁴⁴⁾. Here, we aim to optimize both node and edge conservation while constructing an alignment, as also recognized by a recent effort⁽⁴⁰⁾. In the process, unlike the existing methods that treat each conserved edge the same, we aim to favor conserved edges with NCF-similar end nodes over those with NCF-dissimilar end nodes. And we design WAVE with these goals in mind.

We combine WAVE with NCF of MI-GRAAL as well as with NCF of GHOST. We denote the resulting network aligners as M-W and G-W, respectively. We compare M-W and G-W against the original MI-GRAAL (henceforth also denoted by M-M) and GHOST (henceforth also denoted by G-G), which use MI-GRAAL’s NCF and AS and GHOST’s NCF and AS, respectively. Further, we compare M-W and G-W with a new method introduced recently⁽¹⁰⁾, which is the combination of GHOST’s NCF and MI-GRAAL’s AS (henceforth also denoted by G-M). This allows us to test the performance of WAVE against the performance of MI-GRAAL’s and GHOST’s ASs, under each of MI-GRAAL’s and GHOST’s NCF. We note that we cannot compare M-W and G-W against the combination of MI-GRAAL’s NCF and GHOST’s AS (i.e., M-G), as the current implementation of GHOST does not allow for plugging MI-GRAAL’s NCF into GHOST’s AS⁽¹⁰⁾. Finally, we compare M-W and G-W against the very recent NETAL and MAGNA methods.

We evaluate all methods on synthetic and real-world PPI networks, relying on established data and performance measures^(25,35,24,42,37,44,14). We find that WAVE AS is overall superior to the existing ASs, especially in terms of topological alignment quality. Also, WAVE overall performs comparably to or better than NETAL and MAGNA, especially on synthetic data. This further validates WAVE, because NETAL implements a newer and thus possibly more efficient NCF compared to NCFs of M-W or G-W, which might give NETAL unfair advantage over WAVE.

2 Methods

2.1 Data

We evaluate WAVE on two popular network sets^(25,35,24,42,37,44,14): 1) “synthetic” networks with known node mapping, and 2) real-world networks with unknown node mapping.

The “synthetic” data consists of a high-confidence yeast PPI network⁽⁷⁾ with 1,004 nodes and 8,323 PPIs, and of five noisy networks constructed by adding to the high-confidence network a percentage of low-confidence PPIs from the same data set⁽⁷⁾; we vary the percentage from 5% to 25% in increments of 5%. We align the original high-confidence network to each of the five noisy networks, resulting in five network pairs to be aligned. Since we know the correct node correspondence, we can measure to what extent an aligner correctly reconstructs the correspondence.

The real-world set contains binary (yeast two-hybrid, Y2H) PPI networks of four species: *S. cerevisiae* (yeast/Y), with 3,321 nodes and 8,021 edges, *D. melanogaster* (fly/F), with 7,111 nodes and 23,376 edges, *C. elegans* (worm/W), with 2,582 nodes and 4,322 edges, and *H. sapiens* (hu-

man/H), with 6,167 nodes and 15,940 edges. We align each pair of the networks, resulting in six pairs. If we aimed to predict new biological knowledge, we would have evaluated our method on additional PPIs, such as those obtained via affinity purification followed by mass spectrometry (AP/MS). However, since our main focus is method evaluation, of all PPIs, we focus on binary Y2H PPIs because: 1) they have been argued to be of higher quality than literature-curated PPIs supported by a single publication^(54,19), and 2) the same Y2H networks have already been used in many existing studies^(25,35,24,42,37,44,14). Ultimately, what is important for a fair evaluation is that all methods are tested on the same data, be it Y2H, AP/MS, or other PPIs⁽¹⁴⁾.

When we combine within NCF nodes' topological similarity scores with their sequence similarity scores (see below), for the latter, we rely on BLAST bit-values from the NCBI database⁽¹⁾. When we evaluate biological alignment quality with respect to functional enrichment of the aligned nodes (see below), we rely on Gene Ontology (GO) data⁽¹¹⁾ to evaluate the biological alignment quality. We use the same data versions as in our recent work^(37,14,10).

2.2 Combining Topological and Sequence Information Within NCF

We compute the linear combination of topological node similarity scores s_t and sequence node similarity scores s_s of nodes u and v as: $s(u, v) = \alpha s_t(u, v) + (1 - \alpha) s_s(u, v)$. We vary α from 0.0 to 1.0 in increments of 0.1. We do this for all combinations of MI-GRAAL's, GHOST's, and WAVE's NCFs and ASs. When we compare WAVE to recent NETAL and MAGNA, since current implementations of NETAL and MAGNA do not support inclusion of sequence information, for these methods, we only study topology-based alignments (corresponding to α of 1).

2.3 Evaluation of Alignment Quality

If we align graph $G(V_G, E_G)$ to graph $H(V_H, E_H)$ (where $|E_G| \leq |E_H|$) via an injective function $f : V_G \rightarrow V_H$, let us denote with E'_G this edge set: $E'_G = \{(f(u), f(v)) | u \in V_G, v \in V_G, (u, v) \in E_G\}$. Also, let us denote with E'_H the edge set of the subgraph of H that is induced on nodes from V_H that are images of nodes from V_G . $E'_H = \{(f(u), f(v)) | u \in V_G, v \in V_G, (f(u), f(v)) \in E_H\}$. With these notations in mind, we next define alignment quality measures that we use.

2.3.1 Topological Alignment Quality Measures

Node correctness (NC). Given a known true node mapping (which is typically not available in real-world applications), NC is the percentage of node pairs that are correctly mapped by an alignment. If $f^* : V_G \rightarrow V_H$ is the correct node mapping of G to H and $f : V_G \rightarrow V_H$ is an alignment produced by the aligner, then $NC = \frac{|\{u \in V_G : f^*(u) = f(u)\}|}{|V_G|} \times 100\%$ ⁽²⁵⁾.

Edge Correctness (EC). EC represents the percentage of edges from G , the smaller network (in terms of the number of nodes), which are aligned to edges from H , the larger network⁽²⁵⁾. Formally, $EC = \frac{|E'_G \cap E'_H|}{|E_G|} \times 100\%$, where the numerator is the number of conserved edges.

Induced conserved structure (ICS). ICS is defined as $ICS = \frac{|E'_G \cap E'_H|}{|E'_H|} \times 100\%$. It was introduced because EC fails to penalize for misaligning edges in the larger network, i.e., E'_H , as EC is defined with respect to edges in E_G only⁽⁴²⁾. Hence, ICS accounts for this. However, ICS now fails to penalize for misaligning edges in the smaller network, i.e., E_G , as it is defined with respect to edges in E'_H only. Hence, the following measure, S^3 , was introduced recently to penalize for misaligning edges in both the smaller and the larger network⁽⁴⁴⁾.

Symmetric substructure score (S^3). S^3 is defined as $S^3 = \frac{|E'_G \cap E'_H|}{|E_G| + |E'_H| - |E'_G \cap E'_H|} \times 100\%$ ⁽⁴⁴⁾. Thus, S^3 keeps advantages of both EC and ICS while addressing their drawbacks. And it was already demonstrated to be the superior of the three measures⁽⁴⁴⁾. Thus, we discard EC and ICS measures from further consideration, and instead, we report results for S^3 .

The size of the *largest connected common subgraph (LCCS)*⁽²⁵⁾. In addition to counting aligned edges via S^3 measure, it is important that the aligned edges cluster together to form large, dense, and connected subgraphs, rather than being isolated. In this context, a connected common subgraph (CCS) is defined as a connected subgraph (not necessarily induced) that appears in both networks⁽³⁵⁾. We measure the size of the largest CCS (LCCS) in terms of the number of nodes as well as edges, as defined in the MAGNA paper⁽⁴⁴⁾.

In summary, we focus on NC, S^3 , and LCCS. The larger their values, the better the topological alignment quality.

2.3.2 Biological Alignment Quality Measures

To transfer function from well annotated network regions to poorly unannotated ones, which is the main motivation behind network alignment in computational biology, alignment should be of good biological quality, mapping nodes that perform similar function.

Gene Ontology Enrichment (GO). One could measure GO, the percentage of aligned protein pairs in which the two proteins *share* at least one GO term, out of all aligned protein pairs in which both proteins are annotated with at least one GO term^(44,10). In this case, complete GO annotation data is used, independent of GO evidence code.

Experimental GO (Exp-GO). However, since many GO annotations have been obtained via sequence comparison, and since the aligners use sequence information within their NCF, it is important to test the aligners when considering only GO annotation data with experimental evidence codes. This avoids the circular argument of evaluating alignment quality with respect to the same data that was used to construct the alignments^(25,35,24,44,14,10). Thus, we discard GO measure from further consideration, and instead, we report results for Exp-GO.

In summary, we focus Exp-GO. The larger its value, the better the biological alignment quality.

2.4 Our Methodology

2.4.1 Problem Definition

Existing network alignment methods aim to maximize either node conservation or edge conservation. Further, they treat each conserved edge the same. Here, we aim to simultaneously maximize both node and edge conservation, while favoring conserved edges whose end nodes are highly similar. Given a measure of node conservation (denoted as Node Alignment Quality, NAQ) and a measure of (weighted) edge conservation (denoted as Edge Alignment Quality, EAQ), our goal is to optimize the following expression (denoted as Alignment Quality, AQ):

$$AQ(G, H, f) = \beta_n NAQ(G, H, f) + \beta_e EAQ(G, H, f), \quad (1)$$

where β_n and β_e are parameters used to balance between NAQ and NEQ.

As a proof of concept, we use the following measures as NAQ and EAQ (although any other measure can be used instead). We use the sum of NCF scores over all aligned pairs as our NAQ, which we denote as weighted node conservation (WNC). We design a novel measure of edge conservation as our EAQ, as follows. Similar to EC, ICS, and S^3 , this new measure counts the number of conserved edges, but unlike EC, ICS, or S^3 that treat each conserved edge the same, our new measure weighs each conserved edge by the NCF-based similarity of its end nodes, so that aligning an edge with highly similar end nodes is preferred over aligning an edge with dissimilar end nodes. We denote our new EAQ measure as weighted edge conservation (WEC).

Formally, we define WNC and WEC as follows. Given a pairwise node similarity matrix s with respect to the given NCF, we denote similarity between $u \in V_G$ and $v \in V_H$ in this matrix as s_{uv} . Also, we represent the injection $f : V_G \rightarrow V_H$ as a matrix $y_{|V_G| \times |V_H|}$, where $y_{ij} = 1$ if and only if $f(i) = j$ and $y_{ij} = 0$ otherwise. Thus, the matrix satisfies the following three constraints:

$$y_{ij} \in \{0, 1\}, \quad \forall i \in V_G, \forall j \in V_H; \quad \sum_{l=1}^{|V_H|} y_{il} \leq 1, \quad \forall i \in V_G; \quad \sum_{l=1}^{|V_G|} y_{lj} \leq 1, \quad \forall j \in V_H \quad (2)$$

Then:

$$WNC = \sum_{i \in V_G} \sum_{j \in V_H} y_{ij} s_{ij} \quad (3)$$

To formally define WEC, recall the definitions of EC, ICS, and S³ (Section 2.3.1). All three measures have the same numerator, which we can now rewrite as:

$$|E'_G \cap E'_H| = \frac{1}{2} \sum_{i \in V_G} \sum_{j \in V_H} \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} y_{ij} y_{kl} \quad (4)$$

Here, \mathcal{N}_i denotes the neighborhood of node i , i.e., the set of nodes connected to i . Since each conserved edge will be counted twice, the $\frac{1}{2}$ constant corrects for this.

Now, to leverage the weight of conserved edges by the NCF-based similarity of its end nodes (see above), we define WEC as follows:

$$WEC = \sum_{i \in V_G} \sum_{j \in V_H} \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} y_{ij} y_{kl} s_{kl} \quad (5)$$

With WNC as our NAQ and WEC as our EAQ, formally, our problem is to find a matrix y that satisfies Eqn. 2 and maximizes the following objective function:

$$\begin{aligned} AQ(G, H, y) &= \beta_n NAQ + \beta_e EAQ = \beta_n WNC + \beta_e WEC \\ &= \beta_n \sum_{i \in V_G} \sum_{j \in V_H} y_{ij} s_{ij} + \beta_e \sum_{i \in V_G} \sum_{j \in V_H} \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} y_{ij} y_{kl} s_{kl} \end{aligned} \quad (6)$$

Optimizing the WNC part in Eqn. 6 is solvable in polynomial time (e.g., by using Hungarian algorithm for maximum bipartite weighted matching). However, optimizing the whole function on general graphs is NP-hard. We propose WAVE to solve this problem, while allowing for trade off between node conservation and edge conservation (as the two might not always agree).

2.4.2 Weighted Alignment VotEr (WAVE)

We set $\beta_n = \beta_e = 1$, to equally favor WNC and WEC. Evaluating other combinations of the parameters is of interest but is not of our primary focus. Then, we can rewrite Eqn. 6 as:

$$AQ(G, H, y) = \sum_{(i,j) \in V_G \times V_H} y_{ij} \left(s_{ij} + \sum_{(k,l) \in \mathcal{N}_i \times \mathcal{N}_j} y_{kl} s_{kl} \right) \quad (7)$$

Next, we use set $A = \{(u, v) \mid u \in V_G, v \in V_H, y_{uv} = 1\}$ to denote our alignment, so our objective function has set A as a variable. Then, we use a greedy approach to maximize the objective function, as follows. We start with an empty alignment set A_0 . In each step t , given the current alignment A_{t-1} , we calculate the marginal gain of adding an available node pair (u, v) (in the sense that so far v and u are both unaligned) into A . (For a function $f(S)$ with variable S as a set, the marginal gain of adding an element e into S is defined as $f(S \cup \{e\}) - f(S)$.) That is, we calculate: $AQ(A_{t-1} \cup \{(u, v)\}) - AQ(A_{t-1})$. Then, we align the pair (u^*, v^*) with the highest marginal gain, i.e., $A_t = A_{t-1} \cup \{(u^*, v^*)\}$. To calculate the marginal gain efficiently, we keep the current marginal gain of each node pair and update it in each step. The marginal gain of the node pair (u, v) to AQ is s_{uv} at the beginning (when A is empty, if we align this pair, we can only get s_{uv} in WNC part). In each step, note that if we align two nodes $u \in V_G$ and $v \in V_H$, the side effect is that, in the following steps, when we align another pair of nodes $u' \in \mathcal{N}_u, v' \in \mathcal{N}_v$, both the similarity of (u, v)

and (u', v') will be counted once more by the correctly linked edge, namely, the edge $(u, u') \in E_G$ and $(v, v') \in E_H$. Thus, the marginal gain of (u', v') will be $s_{uv} + s_{u'v'}$ more after (u, v) is aligned.

Intuitively, this process is like voting. When a pair of nodes is aligned, this node pair has a chance to vote for their neighbors: when u and v are aligned, all other node pairs in $\mathcal{N}_u \times \mathcal{N}_v$ receive a weighted vote (with weight $s_{uv} + s_{u'v'}$) from (u, v) , and the weight consists of two parts: 1) the “authority” of the voter, i.e., s_{uv} , 2) the “certainty” of the votee, i.e., $s_{u'v'}$.

The weight for the initial votes of each node pair is the original s_{uv} (which forms the WNC part in the objective function). In every round of WAVE, node pair (u^*, v^*) with the highest vote is aligned, and (u^*, v^*) then vote for all the pairs in $\mathcal{N}_{u^*} \times \mathcal{N}_{v^*}$. The current vote that a node pair gets from its aligned neighbors is the marginal gain to objective function of aligning them.

The WAVE pseudocode is shown the Appendix.

3 Results and Discussion

We evaluate five aligners resulting from mixing and matching NCFs of MI-GRAAL and GHOST with ASs of MI-GRAAL, GHOST, and WAVE: M-M, M-W, G-M, G-G, and G-W (Section 1.3). Also, we evaluate WAVE (the best of M-W and G-W) against NETAL and MAGNA.

By comparing M-M and M-W, we can directly and fairly evaluate ASs of MI-GRAAL and WAVE under MI-GRAAL’s NCF. By comparing G-M, G-G, and G-W, we can directly and fairly evaluate ASs of MI-GRAAL, GHOST, and WAVE under GHOST’s NCF. If WAVE AS produces better alignments compared to the existing methods’ ASs under both of the existing NCFs, this would indicate WAVE’s superiority. If WAVE also produces better alignments compared to NETAL and MAGNA, this would even further demonstrate WAVE’s superiority. However, this is not a strict requirement, as the two new methods either implement both different (newer, and thus possibly superior) NCF than any of M-W and G-W as well as different AS (in case of NETAL), which might give them an unfair advantage, or they work on different principles (in case of MAGNA) and could be thus viewed as complementary to WAVE.

For each combination of network pair, value of α (denoting topological versus sequence information within NCF), and alignment quality measure (Section 2), we do the following.

First, to extract the most out of each source of biological information, it would be beneficial to know how much of new biological knowledge can be uncovered solely from topology before integrating it with other sources of biological information, such as protein sequence information^(25,35,24). Thus, we first compare the different NCF-AS methods on topology-only alignments (corresponding to α of 1 within NCF). Also, since NETAL and MAGNA also produce topology-only alignments, here, we can compare WAVE to these methods.

Second, we examine different contributions of topology versus sequence information in NCF (by varying α), and for each method, we choose the best value of α , i.e., the method’s best alignment. Since current implementations of NETAL and MAGNA do not allow for inclusion of sequence information, here, we cannot compare WAVE to these methods.

For “synthetic” (noisy yeast) networks with known node mapping, we report alignment quality with respect to NC, S³, LCCS, and Exp-GO. For real-world PPI networks of different species with unknown node mapping, we report alignment quality with respect to S³, LCCS, and Exp-GO.

3.1 Comparison of the Five NCF-AS Methods

Here, we compare M-M, M-W, G-M, G-G, and G-W, to test whether WAVE AS improves upon ASs of MI-GRAAL and GHOST under the same (MI-GRAAL’s or GHOST’s) NCF.

3.1.1 Networks With Known Node Mapping

Topological alignments. WAVE is always superior to the existing methods (M-W is superior to M-M, and G-W is superior to G-M and G-G), for all noise levels and alignment quality measures, under both MI-GRAAL’s and GHOST’s NCFs (Figures 1 (a) and 2).

WAVE in general works better under MI-GRAAL’s NCF than under GHOST’s NCF, as M-W is overall superior to G-W. WAVE (at least one of M-W and G-W) beats both MI-GRAAL and GHOST (all of M-M, G-M, and G-G) in 20/20=100% of all cases (Figures 1 (a) and 2).

Best alignments. Under MI-GRAAL’s NCF, WAVE is always superior (M-W is better than M-M), for all noise levels and alignment quality measures (see the Appendix).

Under GHOST’s NCF, WAVE is always superior to MI-GRAAL’s AS (G-W is better than G-M), and WAVE is overall superior to GHOST’s AS (G-W is better than G-G) with respect to two of the four measures (edge-based S^3 and LCCS), while GHOST’s AS is superior (G-G is better than G-W) with respect to the other two measures (node-based NC and Exp-GO) (see the Appendix). Hence, WAVE and GHOST’s AS are comparable overall.

Again, WAVE in general works better under MI-GRAAL’s NCF than under GHOST’s, as M-W is overall superior to G-W. WAVE (at least one of M-W and G-W) beats both MI-GRAAL and GHOST (all of M-M, G-M, and G-G) in 6/10=60% of cases dealing with the two edge-based measures of alignment quality, (see the Appendix).

3.1.2 Networks With Unknown Node Mapping

Topological alignments. Under MI-GRAAL’s NCF, WAVE is always superior (M-W is better than M-M) with respect to S^3 , it is almost always superior with respect to LCCS, and it is sometimes superior with respect to Exp-GO (Figures 1 (b) and 3). Hence, under MI-GRAAL’s NCF, WAVE seems to be favored by topological alignment quality measures.

Under GHOST’s NCF, WAVE is superior to MI-GRAAL’s AS (G-W is better than G-M) in almost all cases, for each of S^3 , LCCS, and Exp-GO (Figures 1 (b) and 3). Also, under GHOST’s NCF, WAVE is overall superior to GHOST’s AS (G-W is better than G-G) with respect to Exp-GO but not with respect to S^3 or LCCS (Figures 1 (b) and 3).

WAVE in general works better under MI-GRAAL’s NCF than under GHOST’s NCF, as M-W is overall superior to G-W. WAVE (at least one of M-W and G-W) beats both MI-GRAAL and GHOST (all of M-M, G-M, and G-G) in 14/18=78% of all cases (Figures 1 (b) and 3).

Best alignments. Under MI-GRAAL’s NCF, WAVE is always superior (M-W is better than M-M) with respect to S^3 , and it is almost always superior with respect to LCCS as well as Exp-GO (see the Appendix). Hence, for best alignments, under MI-GRAAL’s NCF, WAVE is even more superior than for topological alignments only.

Under GHOST’s NCF, WAVE is superior to MI-GRAAL’s AS (as G-W is better than G-M) in most cases for each of S^3 and Exp-GO, and in some cases for LCCS. Also, under GHOST’s NCF, WAVE is overall superior to GHOST’s AS (G-W is better than G-G) with respect to Exp-GO but not with respect to S^3 or LCCS (see the Appendix).

Again, WAVE works better under MI-GRAAL’s NCF than under GHOST’s AS, as M-W is superior to G-W. WAVE (at least one of M-W and G-W) beats both MI-GRAAL and GHOST (all of M-M, G-M, and G-G) in 13/18=72% of all cases (see the Appendix).

The fact that WAVE in general works better under MI-GRAAL’s NCF than under GHOST’s NCF further adds to our recent finding that MI-GRAAL’s NCF is superior to other NCFs^(37,14,10).

3.2 Comparison of WAVE with Very Recent Methods

Here, we compare WAVE (the best of M-W and G-W) with NETAL and MAGNA, which became available close to completion of our study. As such, we were unable to include novelties of these methods (and especially NETAL’s NCF) into our methodology. Recall that we compare the three methods on topology-only alignments, for reasons discussed in Section 2.2.

3.2.1 Networks With Known Node Mapping

WAVE is always superior to both NETAL and MAGNA, for all noise levels and alignment quality measures (Figures 1 (c) and 4). Only in two out of 20 cases, MAGNA is superior: with respect to S^3 for two largest noise levels. But this is not surprising, as MAGNA optimizes S^3 .

3.2.2 Networks With Unknown Node Mapping

WAVE is always superior to MAGNA, for all noise levels and alignment quality measures (see the Appendix). Only in one out of 18 cases, MAGNA is superior to WAVE: with respect to S^3 for one of the six network pairs. NETAL is overall superior to the other two methods, especially with respect to topological alignment quality measures (S^3 and LCCS) (see the Appendix). This could be because NETAL has both different NCF and AS compared to WAVE, and as such, its superiority might be a consequence not of its ASs but rather of its NCF. So, if its NCF was fed into WAVE AS, this could perhaps result in a superior new method. This possibility of designing a novel superior method simply by mixing NCF of one method and AS of another method has already been confirmed on several occasions^(37,14).

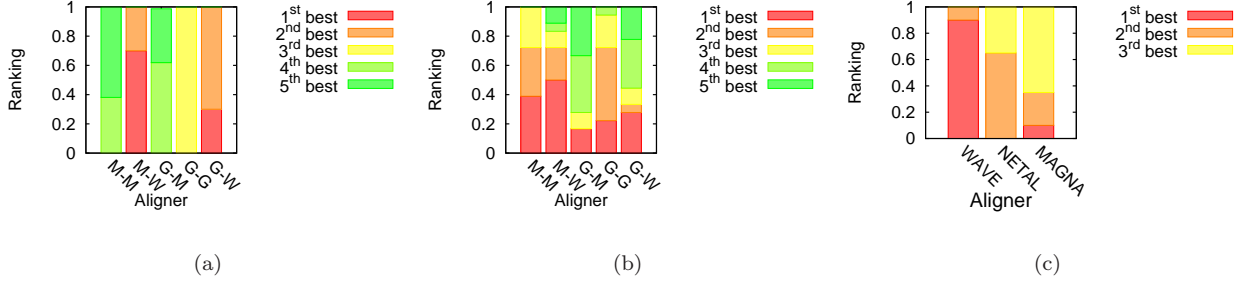


Fig. 1 Representative results for overall ranking of each method over all network pairs in a given data set and over all alignment quality measures. The ranking is expressed as a percentage of all cases in which the given method ranks as the k^{th} best method. (a) Results for the five NCF-AS methods on topology-only alignments of “synthetic” (noisy yeast PPI) networks. For equivalent results for best alignments, see the Appendix. (b) Results for the five NCF-AS methods on topology-only alignments of real-world PPI networks of different species. For equivalent results for best alignments, see the Appendix. (c) Results for WAVE (the best of M-W and G-W) against the recent methods (NETAL and MAGNA) on topology-only alignments of “synthetic” (noisy yeast PPI) networks. For equivalent results for real-world PPI networks, see the Appendix. Details (per network pair and alignment quality measure) for panels (a)-(c) are shown in Figures 2-4, respectively. Recall that M-M and G-G are MI-GRAAL and GHOST.

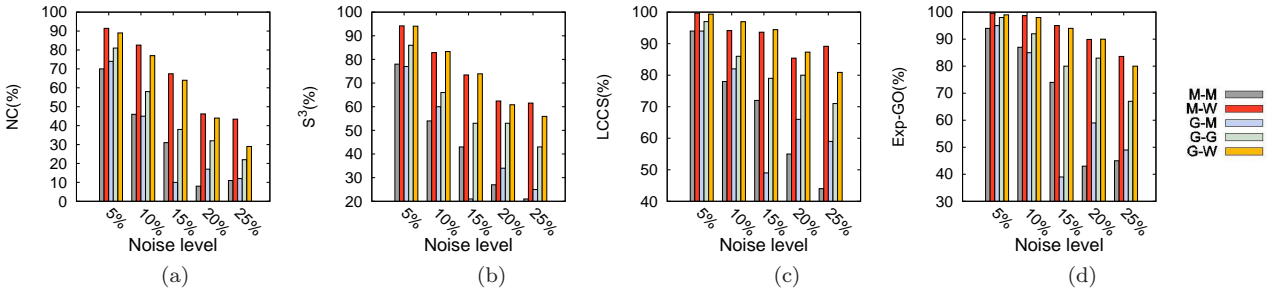


Fig. 2 Comparison of the five NCF-AS methods on topology-only alignments of “synthetic” (noisy yeast) networks with respect to: (a) NC, (b) S^3 , (c) LCCS, and (d) Exp-GO.

4 Concluding remarks

We have presented WAVE, a general network alignment strategy for simultaneously optimizing both node conservation and weighted edge conservation, which can be used with any node cost function or combination of multiple node cost functions. We have demonstrated overall superiority of WAVE against existing state-of-the-art alignment strategies under multiple node cost functions, especially with respect to topological alignment quality. Moreover, we have demonstrated that

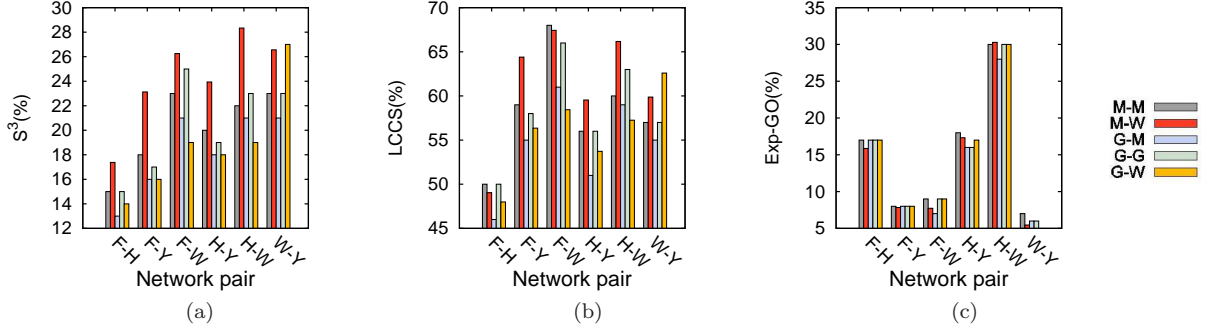


Fig. 3 Comparison of the five NCF-AS methods on topology-only alignments of real-world PPI networks with respect to: (a) S^3 , (b) LCCS, and (c) Exp-GO.

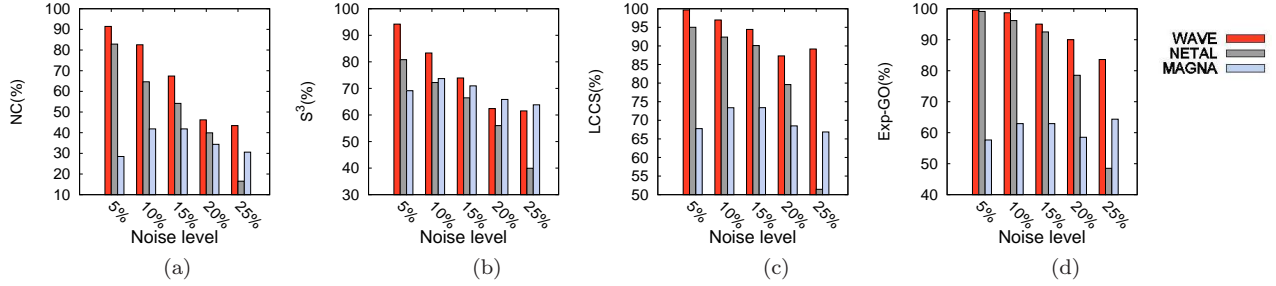


Fig. 4 Comparison of WAVE (the best of M-W and G-W) with very recent network alignment methods on topology-only alignments of “synthetic” (noisy yeast) networks with respect to: (a) NC, (b) S^3 , (c) LCCS, and (d) Exp-GO.

WAVE is comparable or superior even to very recent approaches that became available only close to completion of our study, especially on the sythetic network data. This only further validates the effectiveness of WAVE.

Since WAVE can be combined with any node cost function, doing so for any recent function might improve its alignment quality. Also, WAVE itself can be modified to optimize any other measure of node and edge conservation, which could further improve its accuracy; the measures that we have used are merely a proof of concept that optimizing both node and weighted edge conservation can lead to better alignments compared to optimizing just node conservation (as e.g., MI-GRAAL and GHOST do) or just unweighted edge conservation (as e.g., MAGNA does).

As more biological network data are becoming available, network alignment will only continue to gain importance in the computationally biology domain^(45,6,44). Further, network alignment has implications in many domains. For example, it can be used to de-anonymize online social networks and thus impact privacy⁽³⁹⁾. Hence, further theoretical improvements that would lead to better network alignments have a potential to lead to important discoveries in different fields.

Acknowledgements

Funding: the National Science Foundation CCF-1319469 and EAGER CCF-1243295 grants.

References

- [1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- [2] Mohsen Bayati, Margot Gerritsen, David F Gleich, Amin Saberi, and Ying Wang. Algorithms for large, sparse network alignment problems. In *Data Mining, 2009. ICDM’09. Ninth IEEE International Conference on*, pages 705–710. IEEE, 2009.
- [3] J. Berg and M. Lassig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences*, 101(41):14689–14694, 2004.
- [4] J. Berg and M. Lassig. Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences*, 103(29):10967–10972, 2006.
- [5] Giovanni Ciriello, Marco Mina, Pietro H. Guzzi, Mario Cannataro, and Concettina Guerra. AlignNemo: A Local Network Alignment Method to Integrate Homology and Topology. *PLOS ONE*, 7(6), 2012.
- [6] Connor Clark and Jugal Kalita. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*, page btu307, 2014.
- [7] S.R. Collins, P. Kemmeren, X.C. Zhao, J.F. Greenblatt, F. Spencer, F.C.P. Holstege, J.S. Weissman, and N.J. Krogan. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Molecular Cell Proteomics*, 6(3): 439–450, 2007.
- [8] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.
- [9] S.A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing*, pages 151–158, 1971.
- [10] J. Crawford and T. Sun, Y. Milenković. Fair evaluation of global network aligners. 2014. arXiv:1407.4824 [q-bio.MN].
- [11] João Pedro De Magalhães, Arie Budovsky, Gilad Lehmann, Joana Costa, Yang Li, Vadim Fraifeld, and George M Church. The human ageing genomic resources: online databases and tools for biogerontologists. *Aging cell*, 8(1):65–72, 2009.
- [12] Olivier Duchenne, Francis Bach, In-So Kweon, and Jean Ponce. A tensor-based algorithm for high-order graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2383–2395, 2011.
- [13] Mohammed El-Kebir, Jaap Heringa, and Gunnar W Klau. Lagrangian relaxation applied to sparse global network alignment. In *Pattern Recognition in Bioinformatics*, pages 225–236. Springer, 2011.
- [14] F. Faisal, H. Zhao, and T. Milenković. Global network alignment in the context of aging. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP(99), 2014. ISSN 1545-5963.
- [15] Fazle E Faisal and Tijana Milenković. Dynamic networks reveal key players in aging. *Bioinformatics*, 30(12):1721–1729, 2014.
- [16] J. Flannick, A. Novak, S.S. Balaji, H.M. Harley, and S. Batzoglou. Graemlin general and robust alignment of multiple large interaction networks. *Genome Research*, 16(9):1169–1181, 2006.
- [17] Jason Flannick, Antal Novak, Chuong B. Do, Balaji S. Srinivasan, and Serafim Batzoglou. Automatic parameter learning for multiple network alignment. In *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology*, pages 214–231, 2008.
- [18] X. Guo and A.J. Hartemink. Domain-oriented edge-based alignment of protein interaction networks. *Bioinformatics*, 25(12):i240–1246, 2009.
- [19] Y. Hulovatyy, R.W. Solava, and T. Milenković. Revealing missing parts of the interactome via link prediction. *PLOS ONE*, 9(3):e90073, 2014.
- [20] Brian P. Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R. Stockwell, and Trey Ideker. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research*, 32:83–88, 2004.
- [21] G.W. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1):S59, 2009.
- [22] Danai Koutra, Hanghang Tong, and David Lubensky. Big-align: Fast bipartite graph alignment. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 389–398. IEEE, 2013.
- [23] M. Koyuturk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama. Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, 13(2), 2006.
- [24] Oleksii Kuchaiev and Nataša Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.
- [25] Oleksii Kuchaiev, Tijana Milenković, Vesna Memišević, Wayne Hayes, and Nataša Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of The Royal Society Interface*, 7(50):1341–1354, 2010.
- [26] Simon Lacoste-Julien, Konstantina Palla, Alex Davies, Gjergji Kasneci, Thore Graepel, and Zoubin Ghahramani. Sigma: Simple greedy matching for aligning large knowledge bases. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 572–580. ACM, 2013.
- [27] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. Rimom: A dynamic multistrategy ontology alignment framework. *Knowledge and Data Engineering, IEEE Transactions on*, 21(8):1218–1232, 2009.
- [28] Zhi Liang, Meng Xu, Maikun Teng, and Liwen Niu. NetAlign: A web-based tool for comparison of protein interaction networks. *Bioinformatics*, 22(17):2175–2177, 2006.
- [29] C. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. IsoRankN: Spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–258, 2009.
- [30] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proc. 18th ICDE Conf.*, 2002.

- [31] V. Memišević, T. Milenković, and N. Pržulj. Complementarity of network and sequence information in homologous proteins. *Journal of Integrative Bioinformatics*, 7(3):135, 2010.
- [32] T. Milenković and N. Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008.
- [33] T. Milenković, J. Lai, and N. Pržulj. GraphCrunch: a tool for large network analyses. *BMC Bioinformatics*, 9(70), 2008.
- [34] T. Milenković, V. Memisević, A. K. Ganesan, and N. Pržulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related interaction networks. *Journal of the Royal Society Interface*, 7(44):423–437, 2010.
- [35] T. Milenković, W.L. Ng, W. Hayes, and N. Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121–137, 2010.
- [36] T. Milenković, V. Memišević, A. Bonato, and N. Pržulj. Dominating biological networks. *PLOS ONE*, 6(8):e23016, 2011.
- [37] Tijana Milenković, Han Zhao, and Fazle E. Faisal. Global network alignment in the context of aging. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, BCB’13, pages 23:23–23:32. ACM, 2013.
- [38] Marco Mina and Pietro Hiram Guzzi. Improving the robustness of local network alignment: design and extensive assessment of a markov clustering-based approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 99 (PrePrints):1, 2014.
- [39] A. Narayanan, E. Shi, and B.I.P. Rubinstein. Link prediction by de-anonymization: How we won the Kaggle social network challenge. In *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, pages 1825–1834. IEEE, 2011.
- [40] Behnam Neyshabur, Ahmadreza Khadem, Somaye Hashemifar, and Seyed Shahriar Arab. NETAL: a new graph-based method for global alignment of protein-protein interaction networks. *Bioinformatics*, 29(13):1654–1662, 2013.
- [41] Alexandre Noma and Roberto Marcondes Cesar. Sparse representations for efficient shape matching. In *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*, pages 186–192. IEEE, 2010.
- [42] R. Patro and C. Kingsford. Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105–3114, 2012.
- [43] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23:e177–e183, 2007.
- [44] V. Saraph and T. Milenković. MAGNA: Maximizing Accuracy in Global Network Alignment. *Bioinformatics*, page btu409, 2014.
- [45] R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, 2006.
- [46] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences*, 102(6):1974–1979, 2005.
- [47] R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks. In *Proceedings of Pacific Symposium on Biocomputing 13*, pages 303–314, 2008.
- [48] Rohit Singh, Jinbo Xu, and Bonnie Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Research in computational molecular biology*, pages 16–31. Springer, 2007.
- [49] Aaron Smalter, Jun Huan, and Gerald Lushington. Gpm: A graph pattern matching kernel with diffusion for chemical compound classification. In *Bioinformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.
- [50] R.W. Solava, R.P. Michaels, and T. Milenković. Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics*, 18(28):i480–i486, 2012.
- [51] Fabian M Suchanek, Serge Abiteboul, and Pierre Senellart. Paris: Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment*, 5(3):157–168, 2011.
- [52] Lorenzo Torresani, Vladimir Kolmogorov, and Carsten Rother. Feature correspondence via graph matching: Models and global optimization. In *Computer Vision–ECCV 2008*, pages 596–609. Springer, 2008.
- [53] Lorenzo Torresani, Vladimir Kolmogorov, and Carsten Rother. Feature correspondence via graph matching: Models and global optimization. In *Computer Vision–ECCV 2008*, pages 596–609. Springer, 2008.
- [54] K. Venkatesan, J.F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, and *et al.* An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90, 2009.
- [55] M. Zaslavskiy, F. Bach, and J. P. Vert. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–i267, 2009.
- [56] Mikhail Zaslavskiy, Francis Bach, and J-P Vert. A path following algorithm for the graph matching problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2227–2242, 2009.
- [57] Yutao Zhang and Jie Tang. Social network integration: Towards constructing the social graph. *arXiv preprint arXiv:1311.2670 [cs.SI]*, 2013.